National characteristics and variation in Arabic handwriting

Al-Hadhrami, Ahmed A.N, Allen, Mike, Moffatt, Colin and Jones, Allison Elizabeth

Al-Hadhrami, Ahmed A.N, Allen, Mike, Moffatt, Colin and Jones, Allison Elizabeth ORCID: 0000-0002-9677-3950 (2014) National characteristics and variation in Arabic handwriting. Forensic Science International, 247 . pp. 89-96. ISSN 03790738

Forensic Science International 247 (2015) 89–96

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

National characteristics and variation in Arabic handwriting

Ahmed A.N. Al-Hadhrami [a,b,*], Mike Allen [b], Colin Moffatt [b], Allison E. Jones [b]

[a] Royal Oman Police, Muscat, Oman
[b] University of Central Lancashire, Preston, UK

ARTICLE INFO

Article history:
Received 8 August 2014
Received in revised form 27 November 2014
Accepted 3 December 2014
Available online 12 December 2014

Keywords:
Handwriting
Arabic
Questioned documents
National characteristics
Copybook
Classification

ABSTRACT

From each of four Arabic countries; Morocco, Tunisia, Jordan and Oman, 150 participants produced handwriting samples which were examined to assess whether national characteristics were discernible. Ten characters, which have different configurations depending upon their position in the word, along with one short word, were classified into distinguishable forms, and these forms recorded for each handwriting sample. Tests of independence showed that character forms used were not independent of country ($p < 0.001$) for all but one character-position (this was dropped from subsequent analyses). A correspondence analysis ordination plot and analysis of similarity ($R = 0.326$, $p = 0.0002$) showed that whole samples were discernibly grouped by country, and a tree analysis produced a classification which was 71% accurate for the original data and 83% accurate for 80 new handwriting samples that underwent 'blind' classification. When the countries were combined into two regions, North Africa and Middle East, the grouping was more marked. Thus, there appears to be some scope for narrowing down the nationality, and particularly the wider geographical region of an author based upon the character forms they use in Arabic handwriting.

1. Introduction

Forensic handwriting experts view the static image of handwriting and rarely have the opportunity to directly access the detailed dynamics of the handwriting process by obtaining data from handwriting tablets that provide data about pen movement. For this reason, forensic examiners use subjective methods of examination when making their assessment of a piece of handwriting [1].

Handwriting is a complex skill that requires the integration of both cognitive and motor skills [2,3]. This complexity is apparent at two inter-related levels. The general features used by a person in their handwriting, often called class characteristics, are influenced by the style that is taught and acquired typically during the earlier stages of childhood [4]. Individual characteristics are those that are developed by writers themselves as their handwriting style changes in later years [1].

The style taught formally will vary by location and time [5]. Different countries, for example, may use different educational materials upon which to base the teaching of handwriting [6]. One of the main teaching resources is the 'copybook' which describes ways in which handwriting can be produced including the method of construction and its shape and appearance and the copybooks are likely to vary from place to place [7]. Different copybook styles are likely to be reflected by different class characteristics of the people taught using them and hence the possibility of national characteristics in handwriting features.

National characteristics have been studied in a number of places. Turnbull et al. [8] found class characteristics that to a greater or lesser extent were found to occur more in those with a Polish background compared to those with an English background. On a slightly different theme, Muehlberger [5] found a number of features that were more frequently encountered in the handwriting of Hispanic people within the south east of the USA. Similarly, Cheng et al. [9] found that writers from different racial groups (Chinese, Malay and Indian) in Singapore used a number of different class characteristics.

Studies of national handwriting traits have focussed mainly on the Roman script. Arabic ranks as the sixth language of the world in terms of numbers of native speakers and it is the national and official language in 21 Arab States [10]—Fig. 1. These states stretch from North Africa in the West to the Sultanate of Oman in the East and from Sudan in the South to Syria in the North [11]. A decade ago, literacy rates of Arabic ranged from 40% (Mauritania) to 90% (Jordan) and were increasing [12]. Arabic script is cursive, being

* Corresponding author. Ahmed Al Hadhrami, Royal Oman Police, P. O. Box 131, P. Code 115, Madinat Al Sultan Qaboos, Oman.
Tel: +96899436536; Fax: +96824562993.
E-mail addresses: ahmed.alhadhrami@yahoo.com, gaith1994@yahoo.com (Ahmed A.N. Al-Hadhrami).

http://dx.doi.org/10.1016/j.forsciint.2014.12.004
0379-0738/

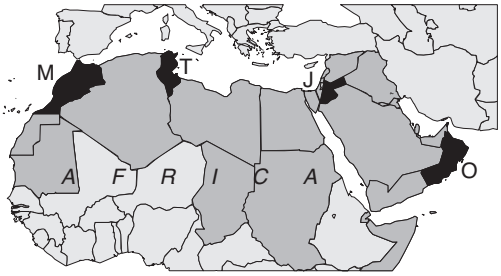

Fig. 1. Location of the countries where samples were collected (in black; M—Morocco, T—Tunisia, J—Jordan, O—Oman). Additional countries with Arabic as an official language are shown in darker grey.

written from right to left and the character forms depend upon whether it is connected as the start, middle or end of a word, or disconnected [13].

Modern Standard Arabic (MSA) is the official means of communication throughout all Arab states. Considerable variation of MSA in handwriting, however, does occur. Arabic characters have different forms according to their position in a word and as it is constructed cursively, there is variation in how characters connect to those preceding or following. The shape, size and relative position of glottal dots (for example ?, ?, and ?) are another source of variation particularly associated with Arabic [14].

In view of the extensive use of written Arabic across a large geographical area, there is a possibility that differing cultural, and particularly educational, factors from one country to another will be apparent in the handwriting from people of different Arab countries. Indeed as handwriting class features gradually change as one moves from one country to another there may be regional variations across the Arab world. This aspect of change across a geographical area has not been studied in detail before.

The primary purpose of this study, therefore, was to determine whether or not Arabic handwriting written in different countries shows class characteristics that indicate the writer's nationality and background, as has been found in studies of Roman script (e.g. [8,9]), or perhaps region. If this were possible, it would have value in a forensic context by potentially providing intelligence about the writer based upon a handwriting sample. A straightforward and efficient way of assessing any potential national or regional characteristics from the static image was required. For this reason, a subjective method of feature extraction was used in this study (as also used for example by Cheng [9]).

2. Methods

From each of four Arabic countries; Morocco, Tunisia, Jordan and Oman (west to east as shown in Fig. 1), 150 participants produced Arabic handwriting samples via a questionnaire which required the writing of particular standard passages. An individual in each country managed the collection of samples to capture demographic and geographic variation in adult native residents in line with ethical processes of the University of Central Lancashire and Royal Oman Police. Countries were chosen based upon where agents were available and the collection of samples seemed feasible, in addition to the desire to include geographic variation.

The choice of participants for obtaining the samples for the analysis was based primarily on two factors, the first being that they be in the age group of eighteen to about seventy and the second factor being that they have attained a sufficient level of basic education. Otherwise, the selection of participants was random from a number of different places to avoid any local effects that might skew the sample.

An even balance of male and female participants was achieved in all countries except Morocco where six more males than females took part. The written passages contained all alphabetic characters in positions unconnected (free), and at the start, middle and end of a word (denoted F, S, M and E, respectively), all written on ruled A4 white paper. Information on each participant's age, gender, handedness and education level was also recorded (Table 1).

Samples were scanned and stored digitally allowing enlargement to visually discern finer detail. In order to analyse handwriting characteristics shared by large numbers of people, it was first necessary to consider what kinds of feature could be analysed. This can only be done manually by a process of feature extraction which involves examining some of the samples of handwriting looking for patterns of feature use that seem to differ between people from different countries. Ten characters were thus chosen to represent diversity in structure, and one short word made of two characters (Lam-Alif). The characters have different configurations depending upon their position in a word, which meant a total of 37 characters/forms were examined. Distinguishable forms of these characters were then identified (Table 2) as all handwriting samples were analysed by the same investigator. The character form used in each handwriting sample was recorded for each of the characters/positions.

For four of the characters (Jeem, Thaa, Ghayn and Kaf) chosen arbitrarily, a second investigator, a non-Arabic speaking individual,

Table 1
Demographic data as frequencies and percentages of the 150 participants from each country.

Variable	Category	Morocco		Tunisia		Jordan		Oman		Overall	
		No	%	No	%	No	%	No	%	No	%
Age	Under 20	18	12.0	9	6.0	18	12.0	4	2.7	49	8.2
	20–29	78	52.0	77	51.3	81	54.0	90	60.0	326	54.3
	30–39	24	16.0	47	31.3	28	18.7	40	26.7	139	23.2
	40–49	21	14.0	8	5.3	16	10.7	10	6.7	55	9.2
	50–59	8	5.3	8	5.3	7	4.7	4	2.7	27	4.5
	60–69	1	0.7	1	0.7	0	0.0	2	1.3	4	0.7
Gender	Female	72	48.0	75	50.0	75	50.0	75	50.0	297	49.5
	Male	78	52.0	75	50.0	75	50.0	75	50.0	303	50.5
Education	Preparatory	9	6.0	14	9.3	11	7.3	7	4.7	41	6.8
	Secondary	88	58.7	46	30.7	40	26.7	43	28.7	217	36.2
	Diploma	10	6.7	2	1.3	36	24.0	23	15.3	71	11.8
	Further	39	26.0	85	56.7	60	40.0	75	50.0	259	43.2
	Without	4	2.7	3	2.0	3	2.0	2	1.3	12	2
Handedness	Both	6	4.0	2	1.3	2	1.3	1	0.7	11	1.8
	Left	5	3.3	17	11.3	10	6.7	13	8.7	45	7.5
	Right	139	92.7	131	87.3	138	92.0	136	90.7	544	90.7

Table 2
Character forms ranked by abundance in 150 handwriting samples from each of Morocco, Tunisia, Jordan and Oman. Up to the six most abundant are shown. Abbreviations are used throughout paper where suffixes refer to position of the character; F—free, S—start, M—middle and E—end of word. Values to the right of illustrated typical form are abundance as percentages (rounded).

			Rank abundance of character forms											
Character	Abbr.	No.	1		2		3		4		5		6	
Ghayn	GF	4		71		21		7		2				
	GS	4		47		26		21		6				
	GM	6		31		31		25		6		6		2
	GE	6		30		30		20		8		7		4
Haa	HF	5		61		24		12		1		1		
	HS	4		62		25		12		2				
	HM	11		17		15		13		10		9		8
	HE	8		48		23		9		5		5		4
Jeem	JF	8		32		19		18		10		1		5
	JS	8		23		21		20		18		12		4
	JM	8		33		28		23		7		5		3
	JE	7		42		33		13		5		4		3
Kaf	KF	6		54		31		11		3		1		
	KS	10		27		17		16		15		9		3
	KM	8		29		23		15		14		9		5
	KE	6		42		25		17		8		5		3
Lam Alif	LA	6		40		14		14		13		13		7
Meem	MF	5		65		13		12		8		2		
	MS	5		34		25		23		9		9		
	MM	6		27		20		19		16		13		6
	ME	6		40		16		16		14		7		7
Thaa	ThF	5		61		22		10		7		1		
	ThS	10		31		21		17		10		7		5
	ThM	5		46		29		15		7		2		
	ThE	6		44		22		21		5		5		3
Thal	TlF	4		53		31		13		3				
	TlE	4		64		21		9		6				
Yaa	YF	4		40		24		22		14				
	YS	4		46		28		19		7				
	YM	4		55		18		17		10				
	YE	4		53		22		16		10				
Za	ZaF	6		41		29		19		7		3		2
	ZaS	7		52		21		14		5		4		3
	ZaM	7		59		18		11		5		4		2
	ZaE	6		57		21		8		8		4		2
Zai	ZiF	4		45		19		18		18				
	ZiE	4		54		25		14		7				

was asked to classify their character forms independently. This was done to establish how objective and transferable the classification procedure was.

2.1. Statistical analysis

For each of the characters tested, a contingency table was constructed of the frequency amongst writers of each identified character form for each of the four countries. Thus, these tables were of size $4 \times n$ where n was the number of forms for that character and had a total frequency of 600. A chi-squared test of independence was then carried out on each of these tables. Where calculated expected values were less than five, the calculated statistic does not have a chi-squared distribution, so a robust method was used to calculate the probability in these cases, where frequencies within the contingency table were randomly allotted while preserving marginal totals. The actual chi-squared value was then compared to a thousand values derived from the randomised allotment to produce a probability of independence [15] for each character form.

Multivariate statistical methods were used to determine whether handwriting samples could be grouped based upon similarity according to the character form attributes recorded. To achieve this, the raw data were first arranged in a matrix such that each row represented a handwriting sample and each column a character form, with 0s and 1s representing absence and presence respectively. Subsequently, correspondence analysis was used to produce an ordination, which was plotted to give a graphical representation of similarity between handwriting samples. Put simply, correspondence analysis begins with the creation of a dissimilarity matrix where each pair of handwriting samples is compared to produce a numerical representation of difference (there are a number of methods for achieving this, the Bray–Curtis method was chosen arbitrarily for this study as it is the default in the software). If these differences are converted to distances, then three handwriting samples can be drawn on a two-dimensional surface where the distance between them is proportional to their dissimilarity. Every additional sample would theoretically require an additional dimension to capture the differences without error. Correspondence analysis seeks to rationalise this number of dimensions by combining attributes that are not too different. The largest amount of variation is shown in the first dimension which becomes the first axis of the plot, and the next largest amount in the second axis, etc. Thus an ordination plot of the first two dimensions shows relative similarity of each handwriting sample in a two dimensional plane, where the axes are mathematical abstractions. While it is possible to plot any pair of dimensions, e.g., the fourth and fifth, they explain a diminishing amount of variation in the samples and are of correspondingly diminishing interest in understanding the data. Axis values often correlate with some quantifiable attribute which may in some way explain the pattern of the plot. The plotted points can be represented by symbols according to their group, e.g., country, and it can be determined whether the groups appear to be distinct by inspecting the plot. What the process does with the handwriting samples is also done simultaneously with the character forms, and these may be plotted on the same plane. There is an association or correspondence between both samples and character forms in where they are plotted, such that a character plotted away from the origin of the graph 'pulls' handwriting samples in its direction by virtue of its inclusion in those samples. This type of analysis is common in plant ecology and more a technical account can be found in Greenacre [16].

A dissimilarity matrix as used for correspondence analysis is also the basis of another procedure used to test statistically whether groups displayed in the ordination are different, by a procedure known as analysis of similarity (ANOSIM). In simple terms, the average intra-group distance between samples is calculated. Then, individuals are reallocated to a group at random, such that these randomised groups have the same number of members as in the 'real' groups, and a new average intra-group distance is calculated. This second step is repeated a great many times (5000 in each case in this paper) and the 'real' average value is compared to the randomly generated values. The proportion of randomly produced values that are smaller than the true value is the probability that the groups are not significantly different.

Finally, a tree analysis was used on the same presence/absence grid to establish a parsimonious procedure for classifying handwriting as belonging to one of the four countries based upon the character forms used. A tree analysis begins with all samples and seeks to split them sequentially into smaller groups using a character form which best separates the countries. By way of testing the accuracy of the resulting classification, it was used to classify a further 80 handwriting samples, 20 from each of the four countries, collected in a similar way as for the main study, but whose nationality was undisclosed to the examiner.

All statistical analyses were carried out using the software R [17] with additional packages used for correspondence analysis [18], ANOSIM [19] and tree analysis [20], as well as producing the map in Fig. 1 [21,22].

3. Results

The second investigator had good levels of agreement with the first for the four characters independently classified as Table 3 shows, despite this second investigator approaching the classification without prior experience of it. This suggests the character forms are relatively distinct and not the subjective opinion of the first investigator.

The results of the tests for association between each of the characters and the four countries are shown in Table 4. All 37 character/positions produced a significant p-value (all were below $p = 0.00135$ which is the significance threshold after Bonferroni's Correction is applied to mitigate for the fact that as more hypothesis tests carried out, significant p-values are more likely to be produced by chance) with the exception of Yaa when starting a word ($\chi^2 = 8.56$, df = 9, $p = 0.48$). This character was, therefore, excluded from multivariate analyses. Table 4 also shows the character forms most often used by each of the countries.

The first two axes of the ordination from the correspondence analysis grouped by the country are shown in Fig. 2. Less useful character forms (closer to the origin) are not plotted to reduce clutter and so make the plot easier to interpret. The first axis of the ordination accounts for 4.1% of variation, and the second axis 1.8%; much of the variation is not reflected in the plot. Even so, there appears to be distinctive grouping, aided by the inclusion of 95% ellipses and standard deviations. Jordan and Oman are most similar and relatively distinct from Morocco and Tunisia with also show much overlap. The result of ANOSIM supported the separation of the four countries ($R = 0.326$, p [5000 randomisations] = 0.0002). Thus there were significant differences amongst

Table 3
Percentage agreement between original investigator and second investigator for character forms when all handwriting samples were analysed independently.

Character	% Agreement
Giim	92
Thaa	91
Ghayn	99
Khaf	93

Table 4
Most abundant of forms examined for each character from handwriting samples from four countries, with percentage of 150 samples present for each country. n is the total number of forms of each character. χ^2 is the result of an association test with $(n-1)\times 3$ degrees of freedom; values marked * had p-value derived by a randomisation method–see Text. All χ^2 values gave $p<0.001$ except that marked † which gave $p>0.05$.

				Country's most common form							
Character	position		n	χ^2	Morocco	%	Tunisia	%	Jordan	%	Oman	%
Ghayn	Free	غ	4	107*	GF1	79	GF1	76	GF1	53	GF1	74
	Start	غـ	4	25.1	GS1	46	GS1	42	GS1	48	GS1	50
	Mid	ـغـ	6	65.3*	GM1	38	GM1	43	GM2	42	GM2	38
	End	ـغ	6	64.2	GE2	44	GE2	34	GE1	39	GE1	33
Haa	Free	ه	5	224*	HF1	37	HF2	40	HF1	97	HF1	82
	Start	هـ	4	241*	HS1	39	HS2	49	HS1	99	HS1	82
	Mid	ـهـ	11	473*	HM2	35	HM2	30	HM8	42	HM7	21
	End	ـه	8	123*	HE1	33	HE1	51	HE1	60	HE1	47
Jeem	Free	ج	8	103*	JF1	31	JF1	28	JF1	27	JF1	41
	Start	جـ	8	206*	JS7	32	JS7	33	JS4	50	JS4	29
	Mid	ـجـ	8	161*	JM3	39	JM1	36	JM1	49	JM2	41
	End	ـج	7	148*	JE1	38	JE1	49	JE1	44	JE2	47
Kaf	Free	ك	6	158*	KF2	49	KF1	42	KF1	70	KF1	65
	Start	كـ	10	97.2*	KS6	36	KS6	31	KS4	26	KS6	25
	Mid	ـكـ	8	80.3*	KM3	37	KM3	31	KM4	30	KM8	25
	End	ـك	6	234	KE1	28	KE5	40	KE1	58	KE1	62
Lam Alif	–	لا	6	393	LA3	45	LA1	47	LA1	58	LA1	47
Meem	Free	م	5	121*	MF1	66	MF1	39	MF1	73	MF1	80
	Start	مـ	5	181	MS2	31	MS3	37	MS2	35	MS2	43
	Mid	ـمـ	6	281	MM6	47	MM6	44	MM5	46	MM2	29
	End	ـم	6	216	ME1	27	ME3	35	ME1	69	ME1	53
Thaa	Free	ث	5	49.5*	ThF1	52	ThF1	58	ThF1	69	ThF1	65
	Start	ثـ	10	117*	ThS3	34	ThS3	30	ThS1	42	ThS1	26
	Mid	ـثـ	5	97.0*	ThM3	52	ThM3	48	ThM2	32	ThM3	53
	End	ـث	6	44.0*	ThE4	43	ThE4	40	ThE4	59	ThE4	35
Thal	Free	ذ	4	88.1*	TlF1	64	TlF1	33	TlF1	49	TlF1	65
	End	ـذ	4	144	TlE1	58	TlE1	42	TlE1	84	TlE1	71
Yaa	Free	ي	4	238	YF2	44	YF2	41	YF1	45	YF1	65
	Start	يـ	4	8.56†	YS2	42	YS2	45	YS2	53	YS2	44
	Mid	ـيـ	4	70.2	YM1	60	YM1	57	YM1	47	YM1	56
	End	ـي	4	61.2	YE2	62	YE2	48	YE2	53	YE2	48
Za	Free	ظ	6	68.4*	ZaF2	35	ZaF2	42	ZaF1	49	ZaF1	46
	Start	ظـ	7	112*	ZaS1	39	ZaS1	55	ZaS1	63	ZaS1	51
	Mid	ـظـ	7	77.5*	ZaM1	41	ZaM1	57	ZaM1	77	ZaM1	61
	End	ـظ	6	178*	ZaE1	35	ZaE1	41	ZaE1	77	ZaE1	74
Zai	Free	ز	4	214	ZiF4	51	ZiF1	58	ZiF2	44	ZiF1	44
	End	ـز	4	57.2	ZiE1	49	ZiE1	56	ZiE1	47	ZiE1	63

the countries in terms of the combinations of character forms used. When region (North Africa = Morocco and Tunisia; Middle East = -Jordan and Oman) was used to produce two groups, the results of the ANOSIM showed even greater group separation by virtue of a higher value of the statistic R ($R = 0.376$, $p < 0.0002$).

Grouping the handwriting samples by recorded attributes other than spatial location brought about significant groupings by ANOSIM for only gender ($R = 0.0095$, $p = 0.003$), the tree analysis of which gave an accuracy of 0.66, which is not too much better than the expected accuracy of 0.5 associated with a random guess. Handedness ($R = -0.004$, $p = 0.55$), age ($R = 0.022$, $p = 0.09$) and education level ($R = -0.001$, $p = 0.51$) as classified on a five point ordinal scale, produced non-significant ANOSIM results.

The results of the tree analysis are shown as a plot in Fig. 3. This plot illustrates the classification of the handwriting samples. For example, of those samples that used a form other than 1 for character HS (Table 2), and used form 4 for character ZiF, 82% were Moroccan writers (47 individuals). The tree can be converted to a series of questions arranged in a dichotomous key. Of the 600 samples, 174 were misclassified by the tree analysis, giving an accuracy of 71%. If countries were again combined into regions of North Africa and Middle East, and another tree analysis performed, 37 samples are misclassified giving an accuracy of 94%. Thus, while the four countries are relatively distinct in terms of the character forms used, regions are much more so.

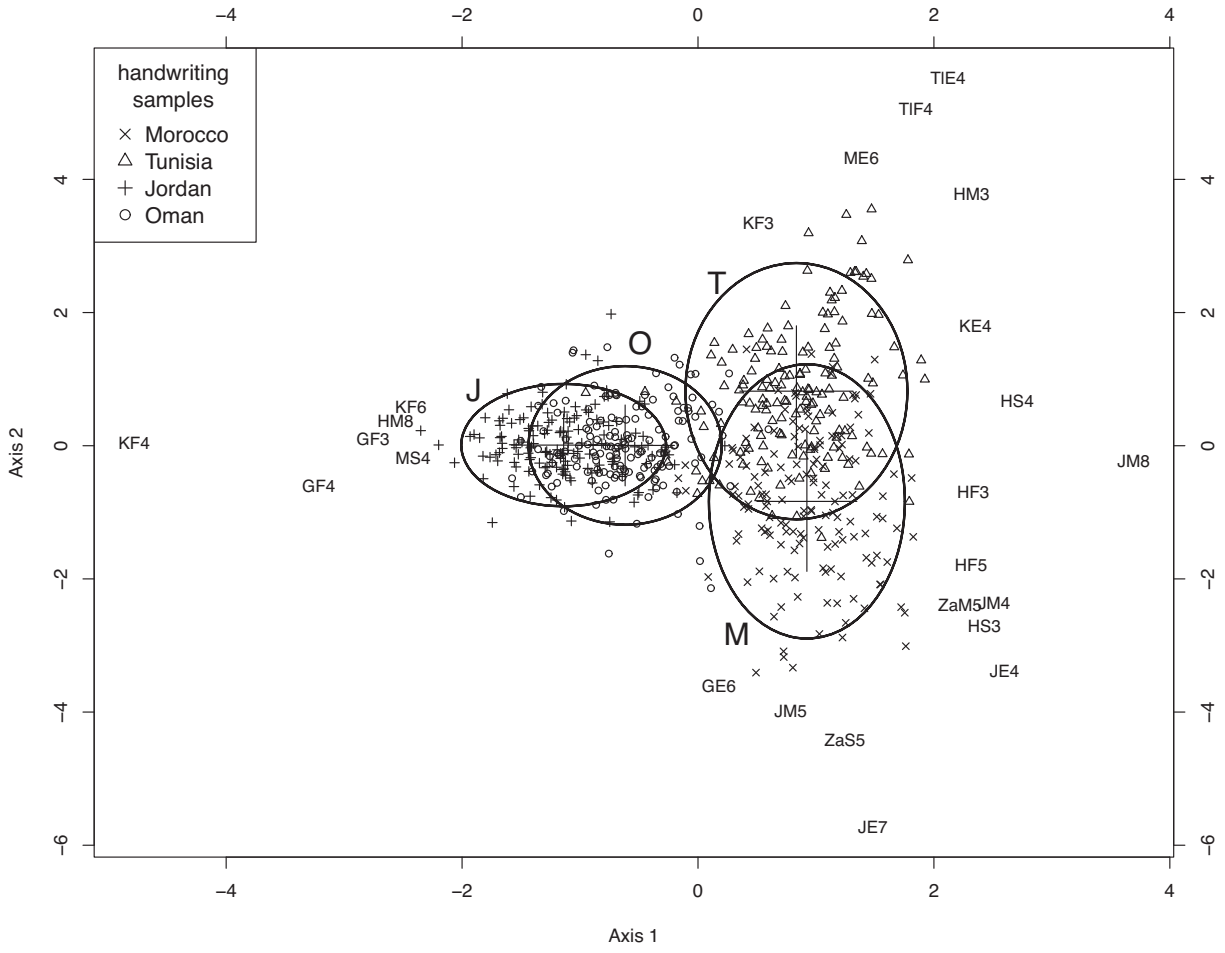

Fig. 2. First two axes of ordination of handwriting sample by correspondence analysis. Ellipses show 95% confidence limit and straight lines show one standard deviation for the samples labelled as M—Morocco, T—Tunisia, J—Jordan and O—Oman. More revealing character forms are also plotted where first letter (or two if second is lower case) correspond to abbreviations given in Table 2, the second upper case letter represents character position (F—free, S—start of word, M—middle of word, E—end of word) and number is character form.

Of the 80 samples (20 from each country) used to test the classification on new data, 66 (83%) were classified correctly (Table 5), with most of the incorrect classifications being within region. When classifying these new samples to region, 76 (95%) were correct.

4. Discussion

4.1. National characteristics

The primary purpose of this study was to test the hypothesis that people with different geographical backgrounds will write Arabic using a number of detectable class characteristics derived from their cultural and educational upbringing. The results clearly provide support for this hypothesis, consistent with findings in which class characteristics have been found in use of the Roman script [8,9].

One possibility that was anticipated was that national style characteristics would be, at least in part, a reflection of the style taught in copybooks. For this reason copybooks were obtained from the four countries in this study and it was found that they were very consistent in their content and the differences subsequently found in this study were not attributable to differences in the taught style.

Since sample size and demographic distributions were similar for the four countries studied, and participation was not based on any selection criteria other than ensuring sufficient literacy to obtain samples in the first place, differences are likely to be attributable to the geographic, and therefore cultural, separation of the sample groups. In this context, none of the four countries (Morocco, Tunisia, Jordan and Oman) are contiguous but rather are separated by significant distances. A likely explanation for the findings of national characteristics is that the cultural (educational) environments in the countries differ such that handwriting styles that are taught or adopted by their populations differ in some ways. Cultural phenomena of this kind are called memes and by their very nature are unpredictable from first principles but rather occur for other reasons such as imitation between people [23].

The process of meme development might be externally influenced too, such as by political influences from outside countries. As travel across the world becomes ever easier, the movement of large numbers of people may also tend to cause the erosion of national handwriting characteristics over time. The findings of this study are consistent with the use of handwriting class characteristics based on the meme model with features being passed from one person to another by a form of (imperfect) imitation. If the imitation were perfect, of course, all writers from a given place would show the same characteristics. The fact that different people in a given place show only some of the features emphasises the individuality of handwriting against this shared cultural background.

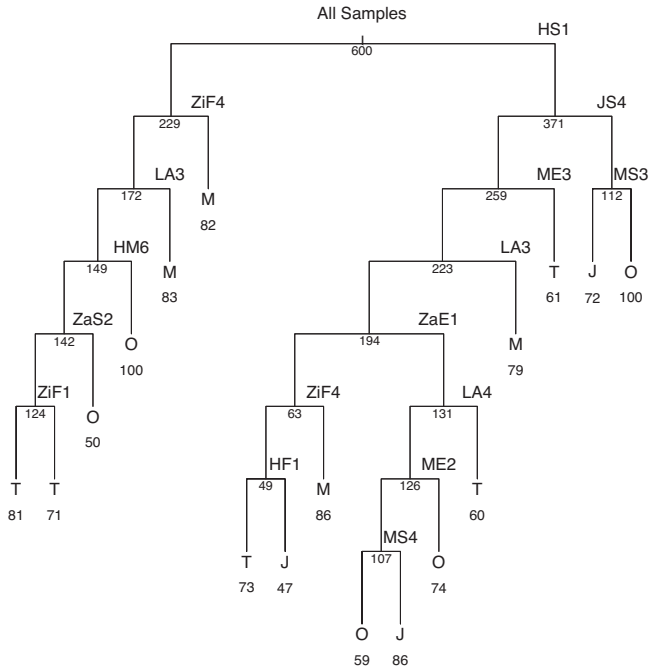

Fig. 3. Tree analysis dichotomous plot for handwriting samples by nationality. Groups are sequentially split into two based upon the most discriminating character form (see Table 2) being present (right branch where character form is shown) or absent (left branch). The abbreviations M, T, J and O show the most abundant sample country of Morocco, Tunisia, Jordan and Oman in the final group (limited to a size of no less than five). Values under the country abbreviation show percentages of the terminal groups which belonged to that country (reflecting classification accuracy) and values under bifurcations show numbers of samples in the group.

Table 5
Classification of new samples (20 from each country) by model created with original data.

		Predicted nationality			
		Morocco	Tunisia	Jordan	Oman
Actual nationality	Morocco	16	1	3	0
	Tunisia	0	19	0	1
	Jordan	4	0	14	2
	Oman	1	2	0	17

4.2. Methodology

The method used to determine which features to use as potential candidates for distinguishing between handwriting from different countries required that samples are examined. The assessment of features and their categorisation is subjective but this is minimised by having clear selection criteria for the categories and reinforced by testing the reliability with more than one rater. There is a danger that the selection of features may become a circular process (looking for features to show that samples of handwriting are different and then using them to show just that). But because it was not possible to predict what features might show national characteristics a priori, this method of feature extraction was the only one possible. This process carries a second danger, namely unintentional cognitive bias in that the rater who is categorising the handwriting sample may attribute features to a piece of handwriting to fit in with some pre-conceived expected outcome, so to remove this as a possibility, the majority of the samples of handwriting were examined blind following an initial phase of feature extraction. Confirmation of the rater categorisation was done by a second rater and there was generally good agreement between the attributed categorisations.

Whilst this study used a particular set of features extracted from the handwriting samples, it is entirely possible that other handwriting features could have been used to show the same result. For this reason, the study is in a sense exploratory demonstrating the principle of finding national characteristics in Arabic handwriting but not excluding the possibility that other features might also be able to show similar patterns of use. Nonetheless, the value of such studies to operational forensic document examiners should also be remembered with the associated need for readily determined features that can impact on an ongoing handwriting investigation.

The manual extraction of features from handwriting can be effectively done by a handwriting expert. An alternative approach is to use image processing to record features from samples of handwriting and to use computer algorithms to extract features which can then be correlated to people from different countries [24]. Using this, success rates for attributing nationality were around 40–50%, somewhat lower than in this study, although more countries were included. Maadeed also reported age and gender differences based on the same features which were more robust than the nationality attributions.

4.3. Regional characteristics

A degree of overlap in handwriting class characteristics might reasonably be expected between adjacent countries (due to closer cultural ties and particularly if borders are porous and people mobile). Such effects might be strengthened by any political or cultural integration between them. In this study, the two North African countries, Morocco and Tunisia, might be expected to share closer ties and likewise the Arabian Peninsula countries, Jordan and Oman, might be expected to be more similar to one another. The findings in this handwriting study show that this is indeed the case with the regional differences between Morocco/Tunisia and Oman/Jordan being even more significant and reliable than the differences between individual countries. It is difficult to provide a definitive cause of this other than to speculate that the cultural phenomena (memes) seem to be shared by those having closer geopolitical ties.

It would be interesting to study more countries so as to fill in the gaps between the four countries studied here with a view to assessing the transition of class handwriting characteristics right across the Arab world. It might also be interesting to see if there are handwriting features that are used in different areas within an Arab-speaking country (similar to [5]).

4.4. Forensic use

The use of national handwriting characteristics to help identify potential authors of a piece of text for forensic intelligence purposes requires the procedure to be robust and reliable enough not to mislead an investigation. Certain handwriting features seem more indicative of nationality than others, but of course it may be that the relevant letters are not present in the piece of handwriting being investigated. The degree of confidence that can be expressed when attributing nationality will, therefore, depend on the features that are present and their value in distinguishing between writers from different countries.

As noted above, whilst the features used in this study showed national and regional differences, it is entirely plausible that other features could be used to provide a more robust and wide-ranging scheme to better distinguish between writers from the four countries in this study and indeed any further countries. There is

also scope for exploring whether handwriting may indicate which region of a country an author is from, which would seemingly be more likely the larger the country and the more socio-political diversity therein.

Acknowledgements

We are most grateful to A.R. Srinivasan for his support and help in the project, to all anonymous participants who supplied the samples and the Royal Oman Police.

References

[1] D. Ellen, Scientific examination of documents: methods and techniques, in: International Forensic Science and Investigation, third ed., CRC Press, New York, NY, 2005.
[2] S. Graham, G.P. van Galen, A review of handwriting research: progress and prospects from 1980 to 1994, Educ. Psychol. Rev. 8 (1996) 7–87.
[3] G.P. van Galen, Handwriting: issues for a psychomotor theory, Hum. Movement Sci. 10 (1991) 165–191.
[4] S. Graham, N. Weintraub, V.W. Berninger, The relationship between handwriting style and speed and legibility, J. Educ. Res. 91 (1998) 290–296.
[5] R.J. Muehlberger, Class characteristics of Hispanic writing in the Southeastern United States, J. Forensic Sci. 34 (1989) 371–376.
[6] R.A. Huber, A.M. Headrick, Handwriting Identification: Facts and Fundamentals, CRC Press, New York, NY, 1999.
[7] R. Sassoon, Handwriting of the Twentieth Century, Routledge, London, 1999.
[8] S.J. Turnbull, A.E. Jones, M. Allen, Identification of the class characteristics in the handwriting of Polish people writing in English, J. Forensic Sci. 55 (2010) 1296–1303.
[9] N. Cheng, G.K. Lee, B.S. Yap, L.T. Lee, S.K. Tan, K.P. Tan, Investigation of class characteristics in English handwriting of the three main racial groups: Chinese, Malay and Indian in Singapore, J. Forensic Sci. 50 (2005) 177–184.
[10] K. Assaleh, T. Shanableh, H. Hajji, Recognition of handwritten Arabic alphabet via hand motion tracking, J. Franklin Inst. 364 (2009) 175–189.
[11] J.R. Smart, Teach Yourself Arabic—A Complete Course for Beginners, Hodder Headline, London, 1992.
[12] N. Bontis, National intellectual capital index: a United Nations initiative for the Arab region, J. Intellect. Cap. 5 (2004) 13–39.
[13] A. Amin, Off-line Arabic character recognition: the state of the art, Pattern Recognit. 31 (1998) 517–530.
[14] M.H. Bakalla, Arabic Culture Through Its Language and Literature, Kegan Paul International, London, 1984.
[15] D.A. Roff, Introduction to Computer—Intensive Methods of Data Analysis in Biology, Cambridge University Press, Cambridge, UK, 2006.
[16] M. Greenacre, Correspondence Analysis in Practice, second ed., CRC Press, Boca Raton, FL, 2007.
[17] R Core Team, R: Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013, , Available from: ⟨http://www.R-project.org/⟩.
[18] O. Nenadic, M. Greenacre, Correspondence analysis in R, with two- and three dimensional graphics: the ca package, J. Stat. Software 20 (2007) 1–13.
[19] J. Oksanen, F.G. Blanchet, R. Kindt, P. Legendre, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, H. Wagner, Vegan: Community Ecology Package, R Package Version 2.0-10, 2013 Available from: ⟨http://CRAN.R-project.org/package=vegan⟩.
[20] B. Ripley, Tree: Classification and Regression Trees. R Package Version 1.0-34, 2013 ⟨http://CRAN.R-project.org/package=tree⟩.
[21] R. Brownrigg, T.P. Minka, Maps: Draw Geographical Maps, R Package Version 2.3-6, 2013 Available from: ⟨http://CRAN.R-project.org/package=maps⟩.
[22] R. Brownrigg, Mapdata: Extra Map Databases, R Package Version 2.2-2, 2013 Available from: ⟨http://CRAN.R-project.org/package=mapdata⟩.
[23] R. Dawkins, The Selfish Gene, second ed., Oxford University Press, Oxford, 1989.
[24] S.A. Maadeed, A. Hassaine, Automatic prediction of age, gender, and nationality in offline handwriting, EURASIP J. Image Video Process. 10 (2014) 1–10 (2014).